CHAPTER EIGHT

The new science of eyewitness memory

Scott D. Gronlund[a,*], **Aaron S. Benjamin**[b]
[a]University of Oklahoma, Norman, OK, United States
[b]University of Illinois at Urbana-Champaign, Champaign, IL, United States
*Corresponding author: e-mail address: sgronlund@ou.edu

Contents

1. Introduction 242
2. Overhauling the old science view 244
   2.1 Relative judgment theory 245
   2.2 Signal detection theory 246
   2.3 Measurement of memory 249
3. Implications of the new science of eyewitness memory 259
   3.1 Re-evaluation of reforms 259
   3.2 Critical role of metacognition 267
   3.3 Balancing costs and benefits 271
4. Conclusions 275
References 277

Abstract

The prevailing view among criminal justice and legal practitioners, and the general public, is that eyewitness evidence is generally inaccurate and unreliable. Here we argue that that perspective fails to take the full cognitive context of eyewitness reports into account. A broader view of *eyewitness cognition* includes both memory judgments—for example, the selection of an individual from a lineup—and an accompanying metacognitive context—for example, the level of confidence that an eyewitness places in that selection. When these components are considered jointly, eyewitness evidence is highly reliable and can be treated like any other source of evidence in the courtroom—valuable when appropriately assayed but prone to contamination.

Empirical research over the past 10 years, based on the bedrock principles of Signal Detection Theory, has illuminated problems with standard historical measures that are based on intuitive theorizing about measurement. Those measures, and the results from experiments that utilize them, have misled the field regarding reform efforts and have diminished the role that eyewitness confidence should play in distinguishing accurate from inaccurate identifications. Signal detection theory, coupled with ROC analysis and confidence calibration, is pointing toward a new science of eyewitness memory. The new science shifts the blame for faulty testimony from unreliable eyewitnesses to other actors in the law enforcement and legal community—actors whose behaviors can

Psychology of Learning and Motivation, Volume 69
ISSN 0079-7421
https://doi.org/10.1016/bs.plm.2018.09.006

transform low-confidence, likely inaccurate, initial identifications, into incorrect, high-confidence, courtroom identifications. Signal detection theory also highlights the role that other metacognitive factors play, as well as how to balance the two types of errors—false identifications of the innocent and missed identifications of the guilty—that inevitably arise from the eyewitness decision problem. The new science of eyewitness memory is leading a transformation in how eyewitness evidence can and should be used by the criminal justice system.

1. Introduction

In the summer of 1984, Jennifer Thompson was raped in her home in Burlington, North Carolina. After fleeing her assailant for a nearby home, another woman in the same neighborhood was raped; the police believed the same man committed both rapes. Jennifer helped with the creation of a composite sketch and a set of suspects was developed and eventually placed into a lineup. After selecting Ronald Cotton from this six-person lineup, Jennifer said, "Yeah. This is the one," adding, "I think this is the guy" (Garrett, 2011b). She later viewed a live lineup, and again chose Cotton, reporting, "This looks the most like him." Largely based on Jennifer's identification, Ronald Cotton was sentenced to life in prison plus 54 years (PBS, 2018).

After serving over 10 years in prison, Cotton was exonerated by the Innocence Project through DNA testing. His case has become a go-to example for the unreliability of eyewitness identification by the courts, the police, and the general public (Thompson-Cannino, Cotton, & Torneo, 2010). This conclusion fit well with the prevailing wisdom of memory researchers, who had demonstrated innumerable ways in which humans were not faithful reporters of past experiences (e.g., Bartlett, 1932; Johnson, Hashtroudi, & Lindsay, 1993; Loftus, 2005; Roediger & McDermott, 2000; Zaragoza & Lane, 1994), and of the myriad factors underlying typical criminal identification procedures that render those reports even less reliable (e.g., Wells & Bradfield, 1998). In fact, understanding the basis of these errors has been an important area of applied psychological research for more than 100 years (Arnold, 1906; Münsterberg, 1908). Honest, well-meaning eyewitnesses, it has long been understood, are unable to reliably identify strangers with much accuracy.

At the core of this belief is an indisputable fact about memory: it is incomplete and prone to error (Schacter, 1999). We forget much of what we experience (Crovitz & Schiffman, 1974; Rubin, 1982); what we do remember, we remember in a manner that is biased by expectations

(Hellmann & Memon, 2016), by our ideas about the way the world works (Pezdek, Finger, & Hodge, 1997), by modifications due to rumination and repeated recounting (Garry & Polaschek, 2000; Loftus & Kaufman, 1992), by confusions with information introduced after the remembered event (Pezdek, 1977), and by the way we are queried (Loftus & Palmer, 1974). It is not surprising, given the wealth and robustness of the evidence on the fragility and malleability of memory, that human memory is seen as an unreliable source of evidence, particularly in the criminal justice system, where the stakes are so high. But this old science view of eyewitness memory is in need of revision.

The old science view is based on an incomplete view both of memory and of the criminal justice system. Memory reports carry *metacognitive* information with them, and that metacognitive information can be useful in eyewitness reports. People can indicate certainty, or lack thereof (e.g., Wixted, Mickes, Clark, Gronlund, & Roediger, 2015; Wixted & Wells, 2017). They can answer a question vaguely or precisely, indicating possession of a lesser or greater amount of relevant knowledge (Koriat & Goldsmith, 1996; Weber & Brewer, 2008). They can say that they *don't know*, indicating an insufficient amount of knowledge to be willing to venture a response (Glucksberg & McCloskey, 1981; Key et al., 2017; Weber & Perfect, 2012). They can respond quickly and confidently, or with hesitation and reluctance (Dunning & Stern, 1994; Sporer, 1992). As we will see below, memory is prone to error but metacognition is generally accurate (Benjamin, 2007; Fiechter, Benjamin, & Unsworth, 2016; Koriat, 2018) and provides useful guidance on the weight that should be applied to a particular memory report.

The differences between the old science and the new science of eyewitness memory owe, at least in part, to unintegrated research traditions (see also Lane & Meissner, 2008). At the beginning of the 21st century, applied research on eyewitness memory and basic research involving recognition memory occurred in different labs, were on divergent paths, and spoke little to one another. In addition, this applied research on eyewitness memory was being conducted at the same time that DNA testing was publicly revealing that many innocent people were in prison, and that faulty eyewitness evidence was one of the reasons why. This was an obvious social ill that memory researchers could help address, and the approach that was adopted was to conduct tests with high ecological validity, and to focus on techniques for improving the quality of typically unreliable eyewitness evidence (what Wells (1978) called *system variable* research). This approach places much of the blame for misidentifications on fundamental limitations of cognition.

Basic research, on the other hand, focused on synthesizing an understanding of core memory functions, including the recognition of previously seen stimuli like faces. Given the many points of common interest between applied and basic memory research, these unintegrated research traditions were unfortunate and to the detriment of both fields. Recent advances at the interface of these fields have promoted a reunification of applied and basic approaches to eyewitness memory (for a review, see Mickes & Gronlund, 2017). In this chapter, we describe the influence of three historically important traditions in basic memory research as they have been imported into eyewitness memory: signal-detection theory, the measurement of memory, and the role of metacognition.

Our chapter begins by reviewing the old science view, and discusses how and why it gained dominion over the field. This will include a review of the theoretical prism through which eyewitness evidence was viewed, and how that prism influenced the ways in which memory was measured and the public reforms that were advocated. Next, we will describe a theoretical framework that offers a different view of eyewitness memory: *Signal Detection Theory* (SDT). SDT has guided basic research in perception and memory for decades (Banks, 1970; Egan, 1975; Green & Swets, 1966; Macmillan & Creelman, 2005). It offers a coherent perspective on decision-making with ambiguous evidence, provides alternative ways of measuring memory, and suggests directions for eyewitness reform efforts. Signal detection theory also provides an explicit linkage to metacognition, and leads to the promotion of confidence as a means of assessing the quality of an eyewitness report. The new science of eyewitness memory is transforming our scientific understanding of eyewitness evidence, and it is beginning to influence the criminal justice system. We will make the case in this chapter that eyewitness evidence is like any other type of forensic evidence: imperfect, but useful when the potential and means for contamination is understood and controlled (Kassin, Dror, & Kukucka, 2013; Wixted & Wells, 2017). Once we consider this new science view, we will revisit the case of Jennifer Thompson and propose a new interpretation for her behavior that conveys a very different message than the simple one that eyewitnesses are unreliable.

2. Overhauling the old science view

To understand the two theoretical perspectives discussed here, and the data marshaled in support of them, it is important to understand general procedures for collecting eyewitness evidence, both in the lab and in the field.

Criminal *lineups* consist of one suspect and several *fillers*—that is, individuals similar to the suspect, but known to be innocent. In the United States, five fillers are usually used, though this varies with convenience and across jurisdictions. Most lineups in the United States are photo lineups, but they can be conducted as live or video presentations (and often are in other countries: Seale-Carlisle & Mickes, 2016). Eyewitnesses are asked to view the lineup and to either select the perpetrator from the set or to indicate that they do not believe the perpetrator to be in that set of photographs. Because fillers are known to be innocent of the crime, they are typically in no legal jeopardy if chosen by an eyewitness. In some cases, only a single photograph of the suspect is used with no fillers; this procedure is called a *showup* and is generally considered to be inferior to a lineup (Neuschatz et al., 2016).

In the field, detectives do not know for sure if the suspect is the real perpetrator of the crime. But, in the laboratory, we can know with certainty if the suspect from a mock crime is included in the lineup or not. So we include *target-present* lineups that have the actual perpetrator, and *target-absent* lineups that have a suspect who is innocent. A target-absent lineup is created by replacing the photograph of the perpetrator with someone else. In some cases, he is replaced by a designated innocent suspect, matched on various physical dimensions. In other cases, another face from the set of fillers used in the rest of the lineup takes the place of the guilty suspect. If an eyewitness selects the guilty suspect from a target-present lineup, this is a correct identification and contributes to the *hit rate*. A false identification occurs when an eyewitness selects the designated innocent suspect from the target-absent lineup, a choice that contributes to the *false-alarm rate*. When a target-absent lineup does not include a single designated innocent suspect, one estimates the rate at which a single suspect would have been chosen by dividing the number of false alarms to all of the fillers by the total number of fillers. The hit rate and false-alarm rate are the basic statistics that describe eyewitness performance in a lineup. After describing the respective theories, we will examine how those statistics are used and interpreted by the competing theories.

2.1 Relative judgment theory

Relative judgment theory (Wells, 1984, 1993) starts from the assumption that there are two bases for reaching the conclusion that a particular member of a lineup is the perpetrator. In a standard lineup, in which the faces are presented simultaneously, eyewitnesses are biased to employ a *relative judgment* rule, in which they compare the members of the lineup to one another, and choose the individual that best matches their memory of the perpetrator. This is not an unreasonable approach to the task if the guilty suspect is actually in the lineup, but it is easy to see how it could lead to an increased likelihood of choosing an innocent suspect from a lineup when the police have the wrong man and he happens to be the best match. Jennifer Thompson appeared to have done exactly that in choosing from the live lineup when she said, "This looks the most like him."

If choosing the best match is all that was involved, eyewitnesses would always choose someone from a lineup; there is always someone who resembles the perpetrator more than the others do. But eyewitnesses do not always choose. This fact makes clear that there is also an *absolute judgment* that is made on the basis of how well a face matches one's memory for the perpetrator, independent of the other members of the lineup. Whereas undue reliance on relative judgments leads to elevated false alarms, it is assumed that a reliance on absolute judgments reduces such biases, and, in so doing, protects innocent suspects. Consequently, according to relative judgment theory, procedures that promote a greater reliance on absolute judgments are sought.

Relative judgment theory has not been formally specified, which makes deriving clear-cut predictions difficult. Also, in part because of the divergence of applied and basic approaches to eyewitness memory, there existed no serious challenger to relative judgment theory for many years. As Bornstein and Penrod (2008) pointed out, a lack of theoretical guidance in the study of eyewitness memory was present from the beginning. Whereas Münsterberg (1908) took an applied approach and made frequent use of examples and anecdotes, Arnold (1906; cited in Bornstein & Penrod, 2008) saw value in theory and was concerned about processes and general principles of memory. Munsterberg's approach carried the day, a historical fact that, we argue, has led to some of the current controversies. Signal detection theory is a prominent current challenger, and the adoption of its theoretical perspective is changing the narrative surrounding eyewitness memory.

2.2 Signal detection theory

Signal detection is a theory of decision-making with wide applicability to tasks involving detection, discrimination, identification, and choice (Green & Swets, 1966). Swets, Dawes, and Monahan (2000) reviewed numerous applications of signal detection theory (SDT), including medical decision-making, predicting violence, detecting cracks in airplane wings, weather forecasting, and law school admissions. What SDT brings to the study of eyewitness memory is a time-tested focus on measurement (Wixted & Mickes, 2012). At the heart of SDT is the idea that the psychological experience of an event is subject to noise (variability), and that that noise can be statistically modeled as arising from a normal distribution. Link (1994; Wixted & Mickes, 2018) traced this history to Fechner's (1860/1966) idea that perception involves "an unknown amount of error that interfered with the measurement of the true value" of a phenomenon of interest. Fechner's theory of discrimination posited that sensations were likewise perturbed by measurement error. He assumed that the decision threshold, or criterion, for discriminating the heaviness of two weights, for example, was located midway between the two Gaussian distributions that summarized the statistical sensory experience of each weight. In the 1950s, the idea of an adjustable decision criterion was adopted (e.g., Egan, 1958), a shift that enhanced the psychological usefulness of signal detection theory and facilitated its adoption into basic research involving recognition memory (Banks, 1970; Lockhart & Murdock, 1970).

We begin with an example of SDT as applied to an eyewitness task; for ease of explication, the example involves a showup (a test with a single candidate face) rather than a lineup, but the extension of the theory to a lineup test is straightforward (see Wixted & Mickes, 2014). The example can be construed as depicting a situation in which a large number of eyewitnesses possess a memory for the face of the perpetrator, but some eyewitnesses had a better view than others. As shown in Fig. 1, a target distribution summarizes the degree to which the test face matches memory for this group of eyewitnesses. A few witnesses had a very good view and formed very strong memories of the perpetrator, and the test face yields a high match to their memories. Others had a very poor view, formed very weak memories, and exhibit lower matches. The memories of the bulk of the eyewitnesses fall somewhere between these two extremes.

Of course, sometimes the police apprehend a suspect who is innocent of a crime. What happens in this case? The suspect may resemble the actual perpetrator to a greater or lesser degree. For idiosyncratic reasons, some eyewitnesses may experience a higher match between this innocent suspect and memory, and others a lower match. But, importantly, the average match across eyewitnesses between memory and an innocent suspect will be lower than between memory and the actual perpetrator. It is for this reason that the distribution of match values for the innocent suspect (the lure distribution) lies to the left of the one for the perpetrator.

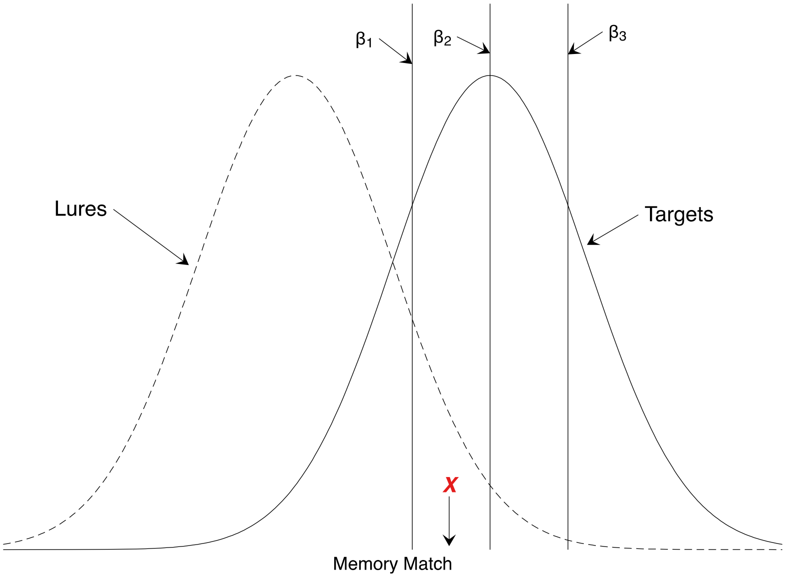

Fig. 1 A depiction of an equal-variance signal-detection model for a showup test. The three levels of instructional bias are denoted by β_1, β_2, and β_3, and *X* denotes the memory strength elicited by a suspect. The hit and false-alarm rates for the criteria located at $\beta_{3,}$ β_2, and β_1 correspond to the data in Table 1 for Conditions A (the careful students), B, and C.

It is critical to note that these target and lure distributions overlap. Consequently, there is no amount of evidence that completely and unambiguously reveals that the match is to the perpetrator or to an innocent suspect. And, in fact, there is a region around the intersection of the distributions where the evidence is highly ambiguous. The degree of overlap of the distributions influences *discriminability*—the latent knowledge differentiating innocent from guilty suspects. If a perpetrator was visible for a longer duration, or eyewitnesses had multiple opportunities to view him, the target distribution would be shifted to the right. This shift would indicate, on average, greater memory strengths for the perpetrator, and consequent greater discriminability from innocent suspects. Alternatively, the variability of the target distribution could be increased (e.g., Ratcliff, Sheu, & Gronlund, 1992), for example, by having eyewitnesses more widely dispersed in space over the crime scene, thereby increasing overlap and decreasing discriminability.

How does SDT conceptualize a decision arising from this framework? The police present an eyewitness with their suspect, and the decision of the eyewitness provides evidence regarding whether the police have the right man. The photo of suspect evokes a memory match in the mind of the eyewitness (denoted by the X in Fig. 1), and the eyewitness compares that match value to a decision criterion. The likelihood that X arises from an innocent or guilty suspect is determined by the nature of the target and lure distributions, the locations and shapes of which are determined by the nature of the memories resulting from the act of witnessing the crime, as well as individual differences associated with that eyewitness. If the match exceeds the decision criterion, the witness selects the suspect; if the match fails to exceed the decision criterion, the witness makes no selection. As indicated above, the position of the decision criterion is adjustable. If a witness is determined that someone must pay for the crime, the value of the decision criterion may take the value β_1, which means that less evidence (a lower memory match) will be needed to prompt a selection. Or the police can inform a witness that the suspect may or may not be present, an act that makes a witness more conservative and shifts the decision criterion t*o* β_2. Or the police may inform a witness that it is important not to implicate an innocent suspect, an instruction that can induce a still higher standard for choosing from the showup, and shift the decision criterion to an even more conservative position (β_3). Critically, within SDT, the position of the decision criterion is both theoretically and empirically separable from discriminability.

2.3 Measurement of memory

Two components of a recognition memory decision, like an eyewitness identification, are crucial to measure. One component involves the accuracy of the decision, so that researchers can determine, for example, whether a proposed reform (for example, sequential lineups) is better than an existing procedure (simultaneous lineups). The second component is the relationship between the confidence expressed by an eyewitness, and the accuracy of that identification (sometimes referred to as eyewitness reliability). We begin with the measurement of accuracy.

2.3.1 Accuracy

No identification or diagnostic technique can be fully evaluated without a joint consideration of both the hit rate and the false-alarm rate that it yields. The theoretical position adopted regarding the origin of these responses dictates the manner of this joint consideration. As we shall see, research on recognition memory research, despite being based on very basic experiments with low ecological validity, was built on a foundation of measurement. The new science of eyewitness memory owes a major debt to this foundation.

Relative judgment theory has much to say about the sources of evidence for endorsement of a face within a lineup. However, it is mostly mute to the important question of how false alarms and hits are related to one another. Consequently, the predominant method by which accuracy was assessed is one that has some face validity but has problematic measurement characteristics. Because hits are desirable, and false alarms are undesirable, Wells and Lindsay (1980) recommended a measure they called the *diagnosticity ratio*: the ratio of the hit rate to the false-alarm rate. This ratio was used to evaluate the quality of techniques for eliciting eyewitness reports. For example, in the first study that compared simultaneous and sequential lineups (Lindsay & Wells, 1985), sequential lineups were deemed superior because the diagnosticity ratio for sequential lineups was greater than for simultaneous lineups. Likewise, in the first study that evaluated the effect of instructing eyewitnesses that the perpetrator may or may not be present in the lineup (Malpass & Devine, 1981), such instructions were determined to be effective because they increased the diagnosticity ratio.

The problem with comparing diagnosticity ratios across conditions owes to the pernicious influence of *response bias*. Response bias can be easily understood by example (adapted from Gronlund, Mickes, Wixted, & Clark, 2015). Take a large class and randomly assign the students into three groups. Random assignment ensures that, on average, these three groups have approximately the same level of knowledge of the course material. Everyone gets exactly the same True-False exam, and everyone is told that they will earn 1 point for each correct response. However, the error payoffs differ between the groups: One group (Condition A) is instructed that each error results in the loss of 10 points, another group (Condition B) is instructed that each error results in the loss of 5 points, and the final group (Condition C) is instructed that each error results in the loss of 1 point. The students in Condition A are going to end up with the fewest correct answers, simply because these students will not answer unless they are absolutely certain that they are correct. Conversely, the students in Condition B and especially Condition C have less to lose and can afford to make some choices with lower confidence. The left two columns in Table 1 provide hypothetical hit rates and false-alarm rates for this example.

Students in Condition A would certainly cry foul if their exam grades were based only on the raw proportion of correct answers—that value is lower for these students simply because they were responding more conservatively,

Table 1 Students randomly assigned into groups that vary in error payoffs on a true-false exam.

	Hit rate	False alarm rate	Accuracy (d')	Response bias (β)	Diagnosticity ratio
Condition A	0.48	0.02	2.0	8.2	24
Condition B	0.60	0.04	2.0	4.5	15
Condition C	0.82	0.14	2.0	1.2	6

Note. Students assigned to Condition A lose 10 points for each error, students in Condition B lose 5 points, and students in Condition C lose 1 point.

a fact that can be easily seen in the lower false-alarm rate. The diagnosticity ratios, shown in the rightmost column, suffer from a similar problem; concluding that the students in Condition A know more than the other students just because they have a larger diagnosticity ratio would be obviously wrong, too.

All that we have done in this example is induce different response biases. Any successful theory of decision-making in a task like this (as for a lineup) must have a means of separating the effects of response bias from the effects of memory sensitivity (the ability to discriminate true from false statements). As we shall see, measures derived from SDT reveal exactly what has happened. The response bias parameter (β) signals that the careful students (Condition A) require a great deal of evidence (i.e., greater certainty) before they are willing to respond, but they have no more knowledge (the d' values are equal across conditions).

Signal detection theory separates accuracy (as assessed by d'), a function of the overlap between the target and lure distributions, from the willingness to make a selection, the response bias (the position of the β's). Fig. 1 depicts how hit and false-alarm rates arise from the underlying evidence distributions. The hit rate is given by the portion of the target distribution that falls above a particular criterion; the false-alarm rate is given by the portion of the lure distribution that falls above that same criterion. Note how the portion of the target and lure distributions that exceed a criterion change together. For example, if the criterion is at β_1, the hit rate will be larger than, say, if the criterion is at β_3, but the false-alarm rate for β_1 also will be larger than if the criterion is at β_3.

A straightforward approach to measuring these two aspects of performance emerges from plotting the hit rate and false-alarm rates at these different levels of response bias. This creates a receiver operating characteristic (ROC) curve. One common way to develop such a curve is to collect responses at varying levels of confidence expressed in the judgment, with the idea that confidence is a proxy for response bias. This means of developing an ROC is the link between SDT and metamemory, described in more detail in the next section.

The assessment of eyewitness performance by constructing ROC curves in this manner is not without its critics (Lampinen, 2016; Wells, Smalarz, & Smith, 2015). Some of the criticism arises from a confusion between theoretical and empirical discriminability (Wixted, Mickes, Wetmore, Gronlund, & Neuschatz, 2017). ROCs can be used as a means of testing competing theories regarding the shape or location of the probability distributions underlying the distribution of evidence. That is a purely theoretical endeavor, and requires the adoption of assumptions that have generally proven correct, but have not been discussed here and are not relevant for the present work. On the other hand, empirical discriminability is a measure of the area under the empirically obtained ROC points and is measured without reference to any theory. Such ROCs are standard in areas like diagnostic medicine and weather forecasting, and are simply a way of combining hit and false-alarm rates into a single, theory-free, entity that reveals underlying discriminability (Swets, 1986).

Returning to the theoretical use of ROCs, any summary measure of performance—including the diagnosticity ratio and d'—makes a prediction about the family of ROC curves that result from manipulating discriminability. The diagnosticity ratio predicts curves that are linear in probability space, whereas d' and related measures predict curves that are bowed. Empirically measured ROCs almost always look like the latter description and almost never like the former, which is how the field has decided that signal-detection theory describes detection and discrimination performance more accurately than other competitors (Rotello, 2017; Swets, 1986).

In the eyewitness domain, we construct the ROC using the choosing rates of guilty suspects (hits) and innocent suspects (false alarms) at different cumulated levels of confidence (Gronlund, Wixted, & Mickes, 2014; Wixted & Mickes, 2012). The top panel of Fig. 2 depicts an ROC curve constructed from the data depicted in Table 1. The only difference is that now β_1, β_2, and β_3 reflect different confidence boundaries (low, medium, and high, respectively) rather than the effects of different error payoffs. The leftmost point on the ROC reflects the proportion of choices of guilty and innocent suspects expressed with high confidence; the middle point on the ROC reflects the correct and false IDs made with either high or medium confidence; and the right-most point reflects the correct and false IDs made

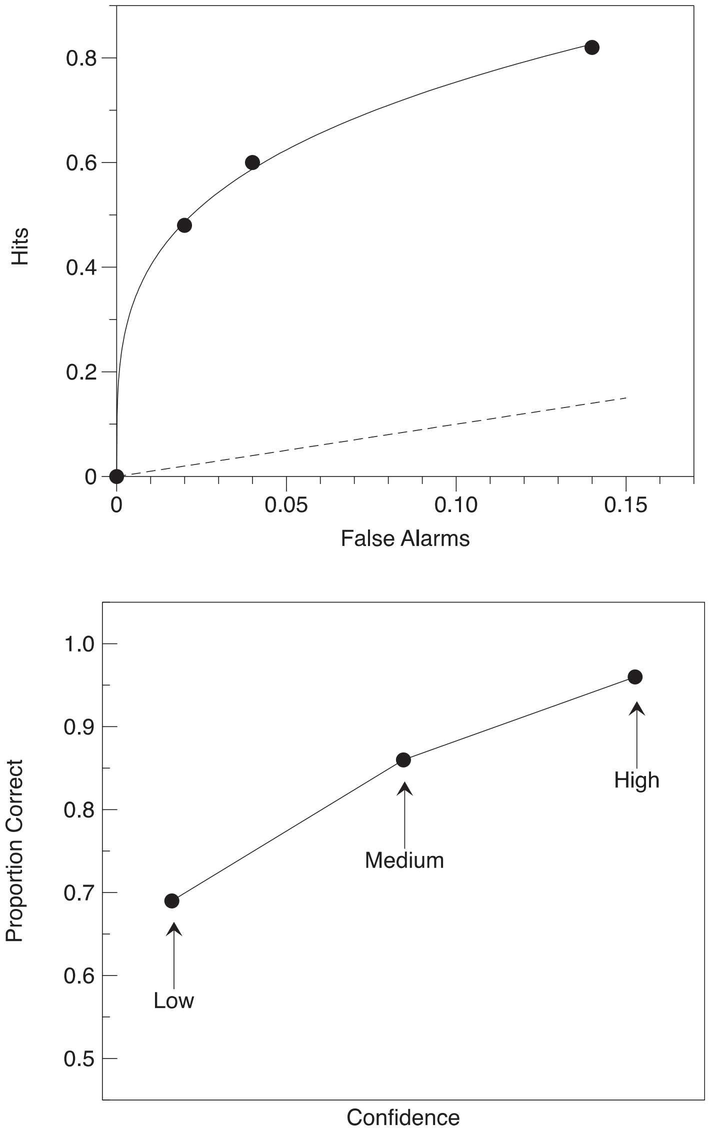

Fig. 2 The top panel depicts an ROC curve based on the signal-detection model depicted in Fig. 1. The high confidence criterion (β_3) results in a correct ID rate of 0.48 and a false ID rate of 0.02; the medium confidence criterion results in a correct ID rate of 0.6 and a false ID rate of 0.04; the low confidence criterion results in a correct ID rate of 0.82 and a false ID rate of 0.14. The dashed line marks chance performance. The bottom panel depicts the calibration curve for these same response proportions. The proportion correct in each confidence category (high: $0.48/(0.48+0.02)=0.96$; medium: $0.12/(0.12+0.02)=0.86$; low: $0.22/(0.22+0.1)=0.69$) is plotted as a function of confidence.

at any level of confidence. The closer an ROC curve falls to the upper left-hand corner of the space, the better discrimination is.

The diagnosticity ratio conflates accuracy and response bias (Wixted & Mickes, 2012), which means that reforms that had been judged as inducing superior accuracy actually may have only been making eyewitnesses less willing to make a selection from a lineup. As experiments have been re-run and analyzed from the perspective of SDT, received wisdom has been upended. It appears as though old-science researchers conflated the empirical aim of determining superior reform procedures with the social aim of reducing the rate at which innocent suspects were being selected by eyewitnesses. Of course, reducing risk to innocent suspects is a commendable goal. But through the prism of signal detection, it is apparent that such actions come with a cost—namely, fewer correct identifications of guilty suspects. An explicit awareness of this unavoidable and ubiquitous tradeoff has contributed to the re-evaluation of the role that eyewitness confidence can play in distinguishing accurate from inaccurate identifications, as we review in the next section.

2.3.2 Confidence

The documentation of techniques that maximize memory discriminability (enhance accuracy) is a core goal of research in eyewitness memory (Clark, Benjamin, Wixted, Mickes, & Gronlund, 2015). But we also need a manner of understanding, and sometimes controlling, the response bias that a decision-maker applies to an identification. In the old science of eyewitness memory, those two goals were conflated. In the new science view, the ROC itself determines diagnostically superior procedures and techniques, and *response confidence* is used to assess response bias. Consequently, measurement of the confidence that accompanies an identification judgment is the second key component of the new science of eyewitness memory.

Confidence has a checkered past in the eyewitness literature. Early work investigating the relationship between confidence and accuracy did so by computing the point-biserial correlation between the two measures. Although one important early meta-analysis reported that correlation to be 0.41 when someone is selected from a lineup (Sporer, Penrod, Read, & Cutler, 1995), the old science view of confidence reports was that they were of little use to the criminal justice system. Penrod and Cutler (1995, p. 830) concluded that eyewitness confidence "… is a weak indicator of eyewitness accuracy even when measured at the time an ID is made and under relatively 'pristine' laboratory conditions." A survey of memory experts (Kassin, Ellsworth, & Smith, 1989; again in

Kassin, Tubb, Hosch, & Memon, 2001) revealed that 87% agreed with the statement, “An eyewitness’s confidence is not a good predictor of his or her identification accuracy.” More recently, Simons and Chabris (2011, p. 5) reported that “...most memory experts agree that an isolated expression of confidence is at best a limited predictor of memory accuracy.” Confidence was thought to be of even less use in circumstances in which memory was deemed generally weak: the *optimality hypothesis* (Deffenbacher, 1980, 2008) suggested that the confidence-accuracy relationship weakens as circumstances for successful memory become less (for example, if the retention interval is longer, if exposure to the perpetrator is shorter, if stress is higher, if a weapon is present). Notably, these are conditions that are common in the witnessing of criminal acts. According to this view, the Supreme Court acted inappropriately when it ruled that eyewitness certainty is one of the factors to be considered when evaluating the accuracy of eyewitness evidence (Manson v. Braithwaite, 1977; Neil v. Biggers, 1972).

Additional reasons to be skeptical about confidence reports came from studies that showed that confidence was malleable as a result of feedback (Wells & Bradfield, 1998, 1999) and repeated remembering (Odinot, Wolters, & Lavender, 2009; Shaw & McClure, 1996). Confidence inflation clearly happened to Jennifer Thompson. The initial confidence statement she offered (“I think this is the guy”) was not deemed sufficient; after she chose Cotton from the live lineup, she asked, “Did I do OK?,” to which the detective responded, “You did great.” Rather than asking her to assess her confidence, the detective asked “if she was certain.” Jennifer Thompson also received reassurances regarding the correctness of her choice from the detective (e.g., “We thought that might be the guy”). But at trial, Jennifer Thompson pointed at Ronald Cotton and reported she was “absolutely sure that Ronald…Cotton is the man” that raped her. Her confidence in that decision had increased massively, but inappropriately, between the original identification and the one she provided in the courtroom.

However, there are important reasons to question the conclusion that confidence is of limited utility in eyewitness reports. To start with, a correlation of 0.41 is not something to be scoffed at within psychological measurement; it signals a medium effect size and is bounded by limited reliability of the component measures. But the concerns go deeper than that, as noted by Juslin, Olsson, and Winman (1996). They showed that the point-biserial correlation can vary from 0 to 1 when individuals are in fact perfectly calibrated—that is, even when confidence exactly matches attained accuracy, the point-biserial correlation can indicate zero relationship between the two. This result obviously indicates that that measure is a poor means of assessing the relationship between confidence and accuracy.

In contrast, SDT provides a natural treatment of confidence that yields a different view of its utility for the criminal justice system. As shown in Fig. 1, memory strength is translated into confidence by extending a set of criteria along the memory match dimension (β_1 through β_3). If the match evoked by a suspect in the showup falls between criteria β_1 and β_2, a witness would express less confidence in that decision than if a suspect's match exceeded β_3. It is clear from Fig. 1 that a greater proportion of the target than the lure distribution falls above β_3 (0.48 vs. 0.02, as shown in Table 1), because a majority of the memory strengths that exceed β_3 arise from the target distribution (the perpetrator), and not the lure distribution (the innocent suspect). Consequently, high confidence will accompany judgments involving strong memory matches, and most of those judgments will arise from events governed by the target distribution. However, this relationship lessens with lower confidence criteria. When considering the lowest criterion (β_1), more similar proportions of values come from the target and lure distributions (0.22 vs. 0.10). Consequently, lower confidence will accompany judgments with weaker memory matches, and those judgments will be less likely to be accurate.

Of course, it is worth remembering what role confidence can and should play in the context of a legal dispute. A judge or jury doesn't want to know about such arcane matters as the point-biserial correlation or the location of decision criteria. Those statistics tell us nothing about the accuracy of a single identification. What is really important is: how accurate one can expect an identification to be, conditional upon a given level of confidence? To answer this question, a second measurement tool, closely related to ROC analysis, is necessary. A calibration curve plots the relative frequency of hits (correct identifications) as a function of confidence. It is straightforward to utilize the data depicted in the top panel of Fig. 2 to construct a calibration curve. The hit and false-alarm rates we need are given in Table 1. The proportion correct for the high confidence judgments (those that exceed β_3) is given by the hit rate (the proportion of the target distribution that exceeds β_3) divided by the sum of the hit rate and false-alarm rate (the proportion of the target and the lure distributions that exceed β_3: $0.48/[0.48+0.02]=0.96$). This value can be computed for each confidence rating. As the bottom panel of Fig. 2 illustrates, as the confidence expressed in an identification increases, the likelihood that the identification is correct increases. (We discuss why this relationship follows from the assumptions of SDT in

Section 3.1.3.) Note that the confidence-accuracy relationship depicted in this manner provides information of direct use to a judge or juror; it indicates how likely decisions expressed at different levels of confidence are to be accurate.

Fig. 3 (reproduced from Wixted & Wells, 2017, figure 5b) shows the strong relationship that exists between confidence and accuracy, collapsed across 15 separate laboratory studies testing lineups. The dashed line on the diagonal depicts perfect calibration, and reveals that laboratory participants clearly are very good at assessing and indicating the likelihood that they are making a correct decision when they make a selection from a lineup (less so when they reject a lineup). It is obvious that, in the controlled setting of the laboratory, the confidence that participants report in their decisions provides information that would be highly informative to the criminal justice system. But does the same relationship hold in the field, when real eyewitness makes consequential decisions from actual lineups constructed and administered by the police?

There are two field studies that are relevant. In one (Klobuchar, Steblay, & Caliguiri, 2006), the researchers measured the frequency and accuracy of rapid identifications that were accompanied by judgments of absolute certainty.

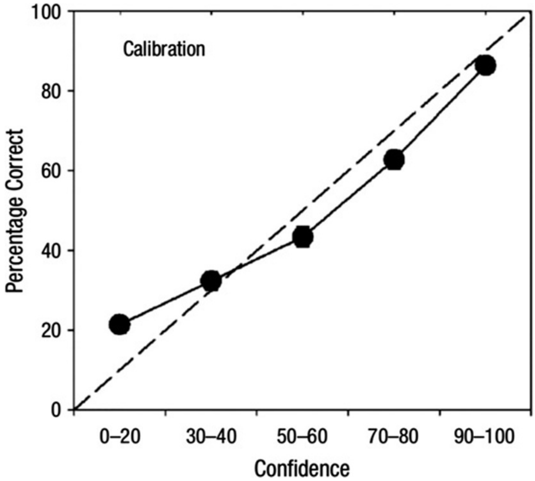

Fig. 3 The calibration curve depicts data from 15 laboratory studies, which all used a 100-point confidence scale. Data are for those participants that made a choice from the lineup. *Reproduced with permission from Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis.* Psychological Science in the Public Interest, 18, *10–65.*

Many of these *jump-out* identifications were made to individuals that the eyewitness knew beforehand and consequently are not of direct interest here. But 26 of these identifications involved strangers, and 25 of those 26 (96%) involved the selection of the police suspect. Of course, it is important to point out that ground truth is not known in the field. That is, unlike in the lab, where an experimenter knows whether a lineup includes a guilty or innocent suspect, in the field the police do not know for certain whether a suspect is guilty or innocent. Yet the fact that suspects were chosen at a much higher rate than matched fillers suggests that these judgments likely exhibited high accuracy.

The second field study (Wixted, Mickes, Dunn, Clark, & Wells, 2016) allowed for a more detailed examination of the contribution of confidence. Eyewitnesses to robberies that took place in the Houston area expressed low, medium, or high confidence in their identifications. They found that the frequency of suspect identifications increased as confidence increased (from less than 20% for low confidence to over 70% for high confidence), and conversely, that the frequency of filler identifications decreased as confidence increased. They estimated the accuracy of the high-confidence suspect identifications to be 97% correct, a value consistent with estimates from laboratory research. They further noted that, as confidence in suspect identifications increased, the proportion of cases increased in which independent corroborating evidence of suspect guilt was also evident. Low-confidence suspect identifications in this study were estimated to be only about 50% correct. These results corroborate the claim that confidence reflects accuracy in eyewitness judgments, even under true crime conditions.

The strong confidence-accuracy relationship bears out the new science view, in contrast to the old science view, which ignores confidence and treats the accuracy of eyewitness evidence as poor. In the next section, we review how the new science of eyewitness memory and the narrative it has inspired have changed ideas regarding reform efforts directed at the collection and treatment of eyewitness evidence. We also discuss how it has brought to the fore two topics that promise to advance understanding of eyewitness memory, and its effectiveness within the criminal justice system: *metacognition*—how people assess their own knowledge, and how they calibrate those assessments to the demands of a particular memory test—and *optimal criterion placement*—how to balance the costs (reduced correct IDs) and benefits (reduced false IDs) inherent in any diagnostic decision involving ambiguous evidence.

3. Implications of the new science of eyewitness memory

The new science perspective fostered by SDT has contributed to a re-interpretation of extant reforms once thought to enhance the accuracy of eyewitness evidence. After reviewing this evidence, we will explore the implications of the *reliability* of eyewitness evidence (the strong accuracy-confidence relationship). We suggest a new narrative in which eyewitness evidence has a key role in distinguishing accurate from inaccurate eyewitnesses when the potential and the means for memory contamination is understood and controlled.

3.1 Re-evaluation of reforms

According to relative judgment theory, the accuracy of eyewitness evidence can be improved in a manner to protect the innocent. To accomplish this goal, the proposed reforms promote absolute judgments, which facilitate "true" recognition (to be contrasted with guesses; Wells, Steblay, & Dysart, 2012). According to the theory, a shift to absolute judgments reduces false identifications but does little to reduce the rate of correct identifications. The initial studies investigating the proposed reforms all yielded data consistent with this "no-cost" view (Clark, 2012). But SDT provides a different perspective on these reform efforts.

A meta-analysis by Clark, Moreland, and Gronlund (2014) evaluated four old science reforms thought to increase the accuracy of eyewitness evidence: *sequential lineups* (Lindsay & Wells, 1985), which force witnesses to view and make a decision about each lineup member individually, *unbiased instructions* (Malpass & Devine, 1981), which instruct the witness explicitly that the perpetrator may or may not be present, *high filler similarity* (Lindsay & Wells, 1980), which dictates that fillers should be similar to the perpetrator, and *matched filler selection* (Wells, Rydell, & Seelau, 1993), which dictates that the fillers should be matched to the eyewitness's description of the perpetrator. In the next two sections, we focus on sequential lineups and unbiased instructions, because interpretation of these two reforms has been most impacted by the new science perspective. For a recent review of issues involving filler selection see Wixted and Wells (2017). As mentioned above, the early studies examining these reforms all showed clear accuracy advantages favoring the reforms. However, all of these effects exhibited decline effects (Ioannidis, 2005; Lehrer, 2010) over the ensuing years, such that, by the time of the Clark et al. meta-analysis, the aggregate accuracy advantages favoring the reforms had entirely disappeared! In fact, instead of increasing accuracy, these reforms simply appear to have made eyewitnesses more conservative responders, and this increasing conservativeness was misinterpreted as increased accuracy due to the reliance on the diagnosticity ratio (Mickes, Flowe, & Wixted, 2012; Rotello & Chen, 2016).

3.1.1 Sequential lineups

Since 2014, a number of laboratory studies (e.g., Andersen, Carlson, Carlson, & Gronlund, 2014; Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Gronlund et al., 2012; Meisters, Diedenhofen, & Musch, 2018; Mickes et al., 2012) and a field study (Amendola & Wixted, 2015; Wixted et al., 2016) have compared sequential to simultaneous lineups using ROC analysis, and the growing consensus is that simultaneous lineups actually *enhance* eyewitness accuracy. It is noteworthy that this is the exact opposite conclusion as that reached by research prior to the use of ROC analysis. As a consequence, the US Department of Justice (Yates, 2017) has withdrawn its recommendation[a] to use sequential lineups.

Why do ROC analysis and diagnosticity-ratio analysis reach different conclusions? It turns out that simultaneous lineups have two effects on memory. To start with, they elicit superior discriminability—that is, it is easier for eyewitnesses to tell the difference between lineups with, and lineups without, a perpetrator. This can be seen in the fact that the ROC for sequential lineups consistently lies closer to the chance diagonal than the ROC for simultaneous lineups. However, sequential lineups also elicit more conservative judgments—that is, they make witnesses less willing to endorse a lineup at all. The diagnosticity ratio confuses this conservatism with greater discriminability, which is why much previous research reached the wrong conclusion about the value of sequential lineups.

The superiority of simultaneous lineups follows naturally from well understood concepts in memory research. Wixted and Mickes (2014) captured these concepts in a signal-detection-based theory that hypothesized why simultaneous lineups result in higher discriminability. By viewing all the lineup members at once, eyewitnesses can determine which characteristics are redundant, or nondiagnostic, and ignore these characteristics.

[a] The recommendation applies to agents at the FBI, Drug Enforcement Administration, Bureau of Alcohol, Tobacco, Firearms and Explosives, the US Marshals Service, and federal prosecutors when deciding whether to charge a case involving an eyewitness identification.

In a fair lineup, all lineup members will share features like race, hair color, and age, so a final decision can't be made on those bases. Constraining the decision through use of a fair, simultaneous, lineup allows eyewitnesses to focus their attention on characteristics that are diagnostic specifically of the perpetrator (e.g., his crooked nose). This specific explanation is not without its critics (e.g., Wells et al., 2015), and other explanations have been proposed (Smith, Wells, Lindsay, & Penrod, 2017; Wetmore, McAdoo, Gronlund, & Neuschatz, 2017), but the general superiority of simultaneous lineups is not in question.

3.1.2 Unbiased instructions

It is recommended that the police inform an eyewitness that the perpetrator may or may not be present in the lineup. The meta-analysis by Clark et al. (2014) showed that these "unbiased" instructions induce a conservative shift in response bias—leading to a higher diagnosticity ratio—but yield no improvement to discriminability (accuracy). A recent study by Mickes, Clark, and Gronlund (2017) and Mickes, Seale-Carlisle, et al. (2017) evaluated directly the effects of different lineup instructions, and compared the outcome of those instructions to an ROC created from confidence ratings.

In her study, two of the instruction conditions induced *liberal* ("better to pick someone…even if you are not sure") or *conservative* ("better to choose the 'not present' option than to pick someone when you are not certain") response biases. The liberal and conservative bias conditions fell slightly below the level traced out by the confidence-based ROC, indicating that such instructions may actually decrease discriminability. Such an effect can be expected if different participants have very different ideas regarding where to place a decision criterion under these conditions (Benjamin, Diaz, & Wee, 2009), or if maintaining confidence criteria imposes a cost on memory (Benjamin, Tullis, & Lee, 2013).

The remaining two instruction conditions, *standard biased* ("If you see the person…please pick him; otherwise, choose the "not present" option") or *unbiased* ("The person…may or may not be in the lineup. If you see the person…please pick him; otherwise, choose the "not present" option"), were modeled after prior research. These two conditions did not differ from one another in discriminability, consistent with the conclusions of Clark et al. (2014). (Surprisingly, they did not differ in response bias either.) But most importantly, the discriminability achieved by these two instruction conditions closely approximated the discriminability achieved in the standard confidence-rating condition.

Generally speaking, as expected by SDT, instructional manipulations and variations in response confidence map out similar relationships, with no evidence that unbiased instructions result in superior accuracy. Given these results, it might seem puzzling that the US National Academy of Sciences, in a recent report reviewing the current state of the field of eyewitness memory (National Research Council, 2014), recommended the use of unbiased instructions. We discuss the rationale for this recommendation in Section 3.3.

3.1.3 Confidence

The role of confidence is the most consequential change to the narrative surrounding eyewitness evidence arising from the new science perspective. Recall from our earlier discussion the optimality hypothesis (Deffenbacher, 1980, 2008)—the claim that, as circumstances worsen for eyewitness memory, the weaker the relationship between confidence and accuracy. This claim is generally inconsistent with research in metacognition: people who have poorer memory—for whatever reason—are reluctant to report those memories and typically do so only in vague terms (Barnes, Nelson, Dunlosky, Mazzoni, & Narens, 1999; Koriat & Goldsmith, 1994, 1996). Eyewitnesses do the same: when asked to report details about a mock crime that they viewed, they withhold low-confidence responses. In one study, information volunteered after a week was as accurate as information reported after 10 min (Evans & Fisher, 2011), revealing the effectiveness with which eyewitnesses withheld information of which they are unsure.

In this section, we review what is known about best practices for soliciting eyewitness identifications and confidence in those identifications. We begin with the assertion that the confidence judgment be made in conjunction with an *initial* test of memory. We then engage in a broader discussion of the *pristine* lineup conditions (Wixted & Wells, 2017) that define the fairness of that initial test. We conclude by revisiting the optimality hypothesis, and describe how SDT explains the way in which a strong relationship between confidence and accuracy is maintained despite factors that adversely affect accuracy.

3.1.3.1 Initial test of memory

The research reviewed in Section 2.3.2 showed that confidence and accuracy are strongly related. However, it is important to emphasize that this is true only if the confidence judgment reflects the initial test of memory (see Steblay & Dysart, 2016). There are two aspects to consider regarding what constitutes an *initial* test. First, it is ideal to assess confidence when it is most diagnostic, which means that the only memory of the suspect that should exist is the one that arises from the original crime event, and not from the eyewitness seeing the suspect in a book of mugshots, in a showup, in the newspaper, or on social media, before viewing a lineup. Second, repeated retrieval attempts can change memory (e.g., Hupbach, Gomez, & Nadel, 2009; Lindsay, 1994). This can happen, for example, if a witness repeatedly describes the perpetrator to multiple detectives before viewing a lineup. Additional research is needed scrutinizing precisely what constitutes an initial test of memory, and is exactly the kind of issue on which expert witnesses can be consulted regarding the individual details of a particular case. In general, pre-identification exposure and repeated retrievals reduce the value of subsequent tests of memory, because they weaken the confidence-accuracy relationship: Interviewers get one good chance to assess an eyewitness's memory and confidence, an idea we will return to at the end of this section in the context of other types of forensic evidence.

3.1.3.2 Pristine lineup conditions

Wixted and Wells (2017) emphasized the importance of collecting eyewitness confidence judgments under "pristine" conditions. Table 2 lists these conditions. In the discussion that follows, Conditions 1 (one suspect per lineup) and 3 (unbiased instructions, the offender may not be in the lineup) primarily impact response bias, whereas Conditions 2 (the suspect should not stand out) and 4 (double-blind testing) are key to defining a fair test of memory. We will deal with these two sets of conditions in turn. Then we turn our attention to Condition 5 (collect a confidence statement at the time of the identification), which is arguably the most important of the five pristine conditions.

Unbiased instructions that inform eyewitnesses that the offender may not be in the lineup (Condition 3) impact response bias, not accuracy (see Clark et al., 2014). But that doesn't mean that including this pristine condition is

Table 2 Pristine lineup conditions from Wixted and Wells (2017).

1. One suspect per lineup
2. The suspect should not stand out
3. Unbiased instructions
4. Double-blind administration
5. Collect the confidence judgment at the time of the identification

superfluous, as long as it is more important to protect the innocent than to implicate the guilty (see Section 3.3). Likewise, Condition 1 (one suspect) is an important safeguard, but it exerts its effects exclusively on response bias: eyewitnesses will be more likely to choose from lineups if they believe that everyone in the lineup is a possible suspect ("One of these people MUST have done it").

Enforcing Conditions 2 and 4 helps ensure that decisions are based on the eyewitness's memory rather than other considerations. Ensuring that no lineup members stand out protects against someone being chosen because he "sticks out" relative to other lineup members (Charman, Wells, & Joy, 2011; Colloff, Wade, & Strange, 2016). Double-blind testing limits any pressure to choose exerted by a lineup administrator. Pressure to choose can take one of two forms. It affects response bias (an eyewitness's willingness to choose *someone*) if a lineup administrator pushes an eyewitness to choose (emboldening eyewitnesses to adopt a more liberal response criterion), although it can influence accuracy if a lineup administrator steers an eyewitness (implicitly or explicitly) toward selecting a particular individual from the lineup (Clark et al., 2015; Clark, Brower, Rosenthal, Hicks, & Moreland, 2013). Although there is some evidence that the absence of double-blind testing actually can *enhance* accuracy in some circumstances (Clark et al. (2013) found it was easier to steer witnesses toward a suspect who is guilty than innocent), sanctioning steering is counter to principles regarding the independence of different types of evidence (e.g., Hasel & Kassin, 2009) and principles of procedural justice (see discussion by Clark, 2012). Eyewitness evidence arising from a lineup in which an eyewitness's selection is swayed by a lineup administrator fails to deliver an independent contribution to subsequent determinations of the guilt (or innocence) of a defendant.

The most important of the five pristine conditions, Condition 5, specifies that a confidence judgment be made in conjunction with the identification decision (and perhaps even video recorded), and not offered at some later time. This is important for several reasons. First, because retrieval itself can modify a memory (e.g., Chan & Lapaglia, 2011; Hupbach et al., 2009; Roediger & Karpicke, 2006), if a confidence judgment is not contemporaneous, the confidence-accuracy relationship can be muddied. Relying upon a confidence judgment made at the time of the identification limits the adverse consequences of confidence inflation (Wells & Bradfield, 1998) and imagination inflation (Garry, Manning, Loftus, & Sherman, 1996), both of which lead to increases in confidence to inappropriate levels over time. Finally, the importance of Condition 5 is reinforced by data showing that an identification need not be fully pristine to be useful. Mickes, Clark, and Gronlund (2017) pointed out that high confidence identifications collected under some non-pristine circumstances (e.g., unfair lineups) remain more likely to signal guilt than low confidence identifications. In other words, confidence can have probative value despite a lack of pristineness, as long as the relied upon confidence judgment is made in conjunction with the identification decision.

The consequences that arise when Condition 5 is violated are severe. The most compelling illustration comes from Garrett's (2011a) analysis of 161 DNA exoneration cases in which faulty eyewitness evidence played a role. It was no surprise to Garrett to find that all of the eyewitnesses were highly confident in their courtroom identifications of the defendants. But for 92 of these cases (57%), Garrett found evidence of what these witnesses reported initially, and *not a single one* of these eyewitnesses expressed high confidence in their initial identifications. (In fact, some had even rejected the initial lineup with the defendant in it, or had chosen an innocent filler.) Among these eyewitnesses was Jennifer Thompson, who initially reported, with considerable hesitation, "I think this is the guy."

The circumstances surrounding these DNA exoneration lineups varied widely (e.g., presence vs. absence of a weapon, same vs. cross-race identifications), as did factors like the length of the retention interval. Yet Garrett (2011a) found that the confidence expressed in these initial identifications was consistently low. If actors in the criminal justice system had relied upon evidence of initial confidence, rather than later, inflated confidence, the eyewitness evidence would have likely played a much smaller role in juror decision-making.

In real criminal cases, circumstances vary and witnessing conditions are often poor. Wixted and Wells (2017; Semmler, Dunn, Mickes, & Wixted, 2018) examined a number of factors that adversely affect identification accuracy, like longer retention interval, shorter exposure to the perpetrator, and greater distance between the eyewitness and the perpetrator. They showed that, although these suboptimal circumstances reduce *average* accuracy, if a high confidence identification is made, it is just as likely to be accurate as the high confidence identifications made under more optimal conditions (e.g., a shorter retention interval). This result again indicates the gatekeeping role that metacognition plays in memory reports. Of course, memory decreases under all of these suboptimal conditions—for that reason, this is a counterintuitive result. How is it that the reliability of the confidence-accuracy relationship is maintained across this wide spectrum of conditions, in contradiction to the optimality hypothesis? Metacognitive assessments of accuracy allow eyewitnesses to sort their memories into confidence categories that more or less maintain accuracy across these conditions. This is a claim that can be more easily understood within the context of SDT.

The finding is consistent with a version of SDT that relies on likelihood ratios to understand how confidence criteria (β_1, β_2, β_3) shift across conditions that affect accuracy (d')—for example, as a function of short vs. long retention intervals. The likelihood ratio is the ratio of the heights of the target and lure distributions at particular locations along the memory-match axis. For example, at the point along the x-axis where the target and lure distributions cross, the likelihood ratio is 1.0: A match value that falls at that point yields totally equivocal evidence for whether it comes from the distribution of targets or the distribution of lures (whether it is a guilty suspect or an innocent suspect). The likelihood ratio increases in magnitude as the criterion becomes more conservative (e.g., β_3) because the height of the target distribution increases as the height of the lure distribution decreases.

Theories of decision-making that use likelihood ratios as criteria assume that participants will naturally maintain relatively constant likelihoods across conditions that differ in accuracy (e.g., Glanzer, Adams, Iverson, & Kim, 1993; Osth, Dennis, & Heathcote, 2017). For example, Stretch and Wixted (1998) found that the confidence criteria fanned out along the memory-match axis for a low- compared to a high-performing discrimination (e.g., short vs. long retention intervals). Within research on eyewitness memory, Semmler et al. (2018) found support for this view when they fit a likelihood-SDT model to data from Lindsay, Semmler, Weber, Brewer, and Lindsay (2008), an experiment that varied the distance between a target person and the eyewitness in a naturalistic setting. Because the likelihood ratios remain relatively constant as a function of factors like distance to a perpetrator, the confidence-accuracy relationship is little changed. Offering an explanation for a surprising finding like this enhances the validity of that finding.

Confidence can provide vitally important information to the criminal justice system about who committed a crime, and often about who didn't. But, to be used fairly and effectively, the police must treat memory like it treats sources of physical evidence. If the police are testing a crime scene for fingerprints, or need to make an impression of a footprint, they cordon off the area to prevent trespass, they wear protective gloves, and they deposit evidence into protective containers. If an eyewitness's memory is treated with the same care, we have shown that the resulting evidence can be highly diagnostic. But just as allowing the public to wantonly traipse through a crime scene likely harms the opportunity to test for fingerprints or to make an impression of a footprint, post-event information and feedback, repeated questioning and testing of an eyewitness, or relying on a delayed confidence judgment, all diminish the value of the eyewitness evidence.

3.2 Critical role of metacognition

The confidence reported in a memory is a key aspect of metacognition, but, as discussed in brief previously, there are other ways in which people calibrate their memory judgments to the demands of a task and the information in their memories. Benjamin (2007) reviews metacognitive influences that affect the encoding and accessing of information, with two processes available to access information, matching and retrieval. Matching arises from the match of a cue to the contents of memory (as is done in a lineup); we already have discussed lineup identifications extensively in this chapter. The other process, retrieval, is characterized by using a cue to extract information related to that cue ("What was the perpetrator wearing?"). There is limited opportunity for an eyewitness to control the encoding of a crime event, given that the event is unexpected, typically of short duration, and often terrifying. Consequently, in this section we focus on the control and monitoring of memory access, especially memory access involving retrieval (see also Fiechter et al., 2016).

3.2.1 Withholding an identification

One decision that is (or should be) under the control of the eyewitness is whether or not to even attempt retrieval. Imagine being asked: "What color was Barack Obama's first tie?" Without even consulting your memory, you can be quite sure that you don't have the relevant information and would be reduced to guessing if forced to answer. Laboratory studies (e.g., Glucksberg & McCloskey, 1981) have shown that participants can quickly and accurately report such an absence of knowledge. More generally, giving people the option to indicate that they don't know an answer allows them to improve the accuracy of the volunteered answers. Weber and Perfect (2012) provide a good example of this effect in an eyewitness situation. They had participants view a mock crime, and make an identification from a showup. Those participants with an explicit "don't know" option performed better (67% of reported decisions were correct) than those participants forced to respond "yes" or "no" (55% correct). Inclusion of a "don't know" option benefits accuracy because eyewitnesses typically know when they do not know something (Bennett, Benjamin, Mistry, & Steyvers, 2018). Of course, this takes place all the time when police canvas a scene for eyewitnesses: If a bystander says she didn't see the perpetrator, her assessment of lack of relevant knowledge is taken to be accurate, and such individuals are rarely asked to view a lineup.

3.2.2 Probing eyewitness memory

How someone decides to search memory for information greatly influences the quality and quantity of retrieved information. Successful rememberers have a plan for how to access information (Indow & Togano, 1970). For example, most US citizens know all 50 states, but reporting them in a haphazard manner likely results in forgetting to report a few, whereas having a plan (e.g., organized by region) will lead to greater success. The same lesson applies to memory reports regarding a crime.

Three approaches are taken by the criminal justice system regarding how eyewitnesses search memory. One approach is to do nothing and leave it up to the eyewitness. This strategy runs the risk that witnesses will choose an ineffective recall strategy and perhaps render critical information even less accessible than before their attempts. A second approach allows the police to guide memory access by posing a series of questions; unfortunately, the ordering of these questions may be haphazard, may include questions that lead an eyewitness toward particular answers (Loftus, 1975), or may lead to interference on important future questions (Shaw, Bjork, & Handal, 1995). Either of these first two approaches can increase the likelihood of inconsistent witness reports due to the critical role of cue dependence on memory success and retrieval failure (Tulving & Thomson, 1973). Retrieval failure is a common cause of forgetting that occurs when cues present at encoding are not present at retrieval, like when you remember a student's name when you see her in class but not when you see her at a local restaurant. Retrieval failures that arise due to recall strategies that change across retrieval attempts contribute to inconsistent witness reports.

Attorneys believe that an inconsistent witness is an inaccurate witness ("If the man that attacked you had a tattoo, why didn't you mention that to the police earlier?"). Metacognition can guide an exploration of this maxim. Stanley and Benjamin (2016) conducted two laboratory studies in which participants studied a set of stimuli followed by multiple free recall attempts. Consistently recalled details were more accurate than inconsistently recalled details—that is, consistently recalled information was more likely to have actually been among the studied material than inconsistently recalled information. But the accuracy of details recalled on later, but not earlier, attempts (*reminiscenced* details, like the tattoo), were just as likely to be accurate as details recalled on earlier but not later attempts (*forgotten* details). Importantly, Stanley and Benjamin also found that, the more inconsistencies in a person's recall, the less accurate the person's consistently produced items were. These findings match prior research using eyewitness paradigms (Krix, Sauerland, Lorei, & Rispens, 2015; Oeberst, 2015), and are in agreement with the behavior of lawyers who rely on inconsistencies in testimony to impeach a witness's credibility.

A third approach to how eyewitnesses search memory involves having eyewitnesses follow a retrieval protocol designed to maximize access to important details and minimize interference. The most prominent example of this approach is the *cognitive interview* (Geiselman et al., 1984; see review by Memon, Meissner, & Fraser, 2010), which imposes a structure on the recall interview, and attempts to limit inconsistencies in witness reports by organizing retrieval attempts in such a way so as to capitalize on cue dependence: Witnesses are instructed to recall from different perspectives (their own vs. others' point of view; Anderson & Pichert, 1978), or in different temporal orders (from the beginning of the event to the end, or from the end of the event to the beginning). A meta-analysis of laboratory studies by Köhnken, Milne, Memon, and Bull (1999) found that the cognitive interview elicited more information than a standard interview, and did so without compromising the accuracy of the reported information. In a field study, Fisher, Geiselman, and Amador (1989) trained experienced detectives in the cognitive interview, and compared the amount and accuracy of the elicited information to a group of equally experienced detectives not trained in the cognitive interview. Replicating the laboratory studies, the trained detectives elicited *more* information from eyewitnesses, with corroboration rates (computed by comparing a witness's report to another reliable source of information) greater than 93% for both sets of detectives. The lesson here is that the use of a technique that treads lightly on memory can improve the quantity of relevant, accurate information that an eyewitness is able to report, and limit inconsistencies in those reports.

3.2.3 Grain size

Once information is retrieved, an eyewitness must decide whether and what to output. Koriat and Goldsmith (1996) proposed that we adjust how much we output in order to achieve an acceptable level of accuracy (Grice, 1975). For example, instructions to be especially accurate reduce the amount of information that rememberers output, and also increase the accuracy of what is output (Koriat & Goldsmith, 1994). This effect is particularly important in the context of memory for conversation, where rememberers have considerable leeway in how detailed a report they provide, and much may depend on exact details of what was said (Brown-Schmidt & Benjamin, 2018; Neisser, 1981).

People adjust the grain size of what they report (Goldsmith, Koriat, & Weinberg-Eliezer, 2002) to achieve acceptable levels of accuracy. That is, people answer with less detailed information when they know (or remember) less, and more detailed information when they know (or remember) more. For example, if asked, "When did the Rolling Stones release their first album?," someone with a lot of relevant information might report 1964, whereas someone with less information might report "in the 60s." Both answers are correct, but are offered at differing levels of detail. Weber and Brewer (2008) had participants view a mock crime video and then answer a series of questions that allowed answers that could vary in grain size (e.g., the duration of the crime, to the nearest 30 s, or within a 2-min range). Participants had to provide both fine- and course-grained answers to each question; participants also reported their confidence in their answers. After completing all of the questions, participants revisited their earlier answers and indicated whether they would nominate the fine- or the coarse-grained answer to provide to a police officer investigating the crime. Not surprisingly, overall accuracy was greater for coarse-grained than fine-grained answers (0.72 vs. 0.44, respectively). However, the accuracy of the nominated answers fell between these two values (0.64), indicating that participants can select an appropriate response on an item-by-item basis, commensurate with the level of expertise they felt they had for that question. Weber and Brewer also showed that the confidence expressed in both fine- and coarse-grained answers was similarly calibrated to accuracy (although there was a general overconfidence expressed in the fine-grained answers). People choose a point that trades off informativeness for accuracy conditional upon their expertise and upon the situational demands.

The critical role that metacognition plays in governing the use of memory suggests new or modified techniques that may enhance the ability to distinguish accurate from inaccurate eyewitnesses, or enhance the quality of the information eyewitnesses provide. Brewer, Weber, Wootton, and Lindsay (2012; Sauer, Brewer, & Weber, 2008) had participants make confidence assessments to each lineup member. They found that classification algorithms (that weigh the pattern of confidence judgments across the set lineup members) sorted accurate from inaccurate eyewitnesses better than did standard binary lineup decisions. Horry, Brewer, and Weber (2016) examined what they called grain-size lineups, in which participants eliminate as few, or as many, lineup members from consideration as they desire (eliminating all but one lineup member is a fine-grained decision). The grain-size lineup is reminiscent of the elimination lineup (Pozzulo & Lindsay, 1999), in which participants must choose one individual from a simultaneous lineup, and then are asked if the selected individual is the perpetrator. Although Horry et al. found that in their most difficult lineup condition, participants produced a greater proportion of coarse-grained responses, neither grain-size lineups, nor elimination lineups, induced better performance than simultaneous lineups.

Researchers should continue to explore new approaches to testing eyewitness memory. Wells, Memon, and Penrod (2006, pp. 68–69) wrote:

> *It could be argued that research has been profoundly conservative in its approach to the eyewitness-identification problem. Specifically, researchers have tended to operate within the confines of the traditional lineup, in which a suspect is placed among fillers and the eyewitness makes a verbal identification. But what if the lineup had never existed and the legal system turned to psychology to determine how information could be extracted from eyewitnesses' memories?...Operating from scratch, it seems likely that modern psychology would have developed radically different ideas.*

Metacognition, and SDT, can help guide modern psychology in this endeavor.

3.3 Balancing costs and benefits

The central claim of the new science perspective is that eyewitnesses, under the proper circumstances, are highly reliable. Notably, this is not the same thing as saying that they are highly accurate. Many are not. But as we have made clear above, they can generally tell us when they are or are not accurate, making them reliable reporters of what they remember. That means that not all identifications should be treated equally; high-confidence judgments, made under pristine conditions, are more diagnostic. But how high should confidence be for the police to treat an identification as truthful? Some might think that the level should be greater than 95%, because there would be very few false identifications of innocent suspects. But, as we have discussed, this will result in fewer correct identifications of guilty suspects, as compared to the adoption of an 85% or 75% confidence level.

3.3.1 Optimal confidence level

What is the ideal confidence level at which to consider eyewitness evidence probative? This is a complicated question, with many factors to consider. Moreover, it is a question that policy makers, not memory researchers, must answer. However, SDT provides guidance for determining the optimal confidence level because it makes explicit the tradeoff between two types of errors: failing to identify a guilty suspect from a lineup (a miss), and the identification of an innocent suspect (a false alarm). This tradeoff is apparent in the reforms discussed above, like sequential lineups, which make eyewitnesses more conservative (reducing false alarms, but at the cost of fewer correct identifications). The same tradeoff is apparent when responses are considered at different levels of confidence, with high confidence judgments resulting in fewer false alarms but more misses (fewer correct identifications), and low confidence judgments producing more false alarms, but fewer misses (more correct identifications). As Clark (2012, p. 248) put it, "What should the exchange rate be for correct identifications lost versus false identifications avoided?"

To answer this question, consider the costs associated with the two errors. How much more important is it to protect the innocent than to let guilty suspects go free? There is a long history concerning the relative costs of these errors. English jurist Sir William Blackstone (1765, p. 352) wrote that it is "…better that ten guilty persons escape than that one innocent suffer"; Benjamin Franklin thought the ratio should be 100:1 (see Baumgartner, Grigg, Ramírez, & Lucy, 2018 for a recent review of this issue).

3.3.2 Role of base rates

Weighing these errors also requires consideration of the frequency with which the police place innocent suspects in lineups. If this happened only rarely, reform efforts that induce more conservative responding are misplaced, protecting against an error that is unlikely to occur. However, policy makers may feel differently if a larger proportion of lineups contain innocent suspects. The base rate with which suspects in lineups are actually guilty is difficult to estimate for a number of reasons. There exist only haphazard records of lineups from which suspects are not chosen, and, of course, conviction doesn't mean that a defendant is guilty. The base rate with which suspects in lineups are guilty also likely depends on when in an investigation a lineup is administered. Wells and Olson (2003) argued that the police should have probable cause before placing a suspect in a lineup, which should reduce the base rate of innocent suspects being put at risk. But in some circumstances, and in some jurisdictions, lineups are used to build a case against a suspect, likely putting many more innocent suspects at risk. Finally, plea bargaining may play a role. The vast majority of criminal cases in the United States are resolved by plea bargain; Rakoff (2014) reports that more than 97% of federal criminal charges that are not dismissed are resolved by plea bargain. A plea bargain can be reached at nearly any stage in the criminal justice process, but if a suspect who is guilty is more likely to take a plea (Gregory, Mowen, & Linder, 1978), plea bargaining becomes a mechanism that may put more innocent suspects at risk in lineups.

Although there exists no empirical estimate of this base rate, Wixted et al. (2016) used SDT to offer a principled, theory-based estimate. Using the Houston field study data, they estimated that only 35% of the lineups constructed by the Houston Police Department included a guilty suspect. More work like this is needed to validate this low estimate. If it is correct, it does suggest that calibration of evidential standards should be more focused on protecting the innocent than on ensuring conviction of the guilty (which is consistent with the impact of unbiased instructions as suggested by the National Research Council (2014) report).

3.3.3 Formalism for balancing costs and benefits

Signal detection theory shows us how to combine the costs of the two types of eyewitness errors with the base rates to determine where a decision criterion should be set (McNicol, 1972; for a detailed treatment, see Green & Swets, 1966). To begin, policy makers must set a positive value for hits (V_{H}) and correct rejections (V_{CR}), and a negative cost for misses (C_{M}) and false alarms (C_{FA}). If it is assumed that the base rate of guilt among suspects in lineups is 50%, the optimal likelihood ratio criterion (β_{optimal}) is straightforward:

$$\beta_{\text{optimal}} = \frac{V_{\mathrm{CR}} + C_{\mathrm{FA}}}{V_{\mathrm{H}} + C_{\mathrm{M}}} \qquad (1)$$

Eq. (2) shows the necessary adjustments if the base rates are not equal, where BR_{G} is the proportion of lineups that include a guilty suspect and $(1 - BR_{\mathrm{G}})$ is the proportion of lineups that include an innocent suspect.

$$\beta_{\text{optimal}} = \frac{(V_{\mathrm{CR}} + C_{\mathrm{FA}}) \times (1 - BR_{\mathrm{G}})}{(V_{\mathrm{H}} + C_{\mathrm{M}}) \times (BR_{\mathrm{G}})} \qquad (2)$$

Now let us consider an example of how to use this equation in practice. Before we do so, remember that the assessment of values and costs is a civic and not a scientific question; we take certain values here only as an example and do not intend this to be a set of recommendations for such values.

Following Blackstone's principle, set the cost of a false alarm to be 10 times that of a miss ($C_{FA} = 10$ vs. $C_M = 1$), and let the benefit of a hit be equal to the benefit of a correct rejection ($V_H = V_{CR} = 1$). If we use the base rate estimate from Wixted et al. (2016, $BR_G = 0.35$), the optimal criterion position is quite conservative, $\beta_{\text{optimal}} = 10.2$. This means that the criterion will be ideally placed at a point where the eyewitness evidence is slightly more than 10 × in favor of guilt than innocence. This conservatism makes sense because more of the suspects are innocent, and the cost of a false alarm is much greater than the cost of a miss, making the goal of reducing false alarms more important than limiting misses. The value of β_{optimal} dictates the level of confidence upon which the police should rely. In the case of such a high prescribed value for β_{optimal}, only identifications at high levels of confidence would be considered probative.

What if, instead, policy makers decide that the two errors should be treated equivalently ($C_{FA} = C_M = 1$), but that the value of selecting guilty suspects is five times more beneficial than rejecting lineups with innocent suspects ($V_H = 5$, $V_{CR} = 1$)? For the same estimate of the base rate of guilt among suspects, the optimal criterion position is more liberal ($\beta_{\text{optimal}} = 0.62$), and lower confidence identifications of suspects would be considered probative.

There are many factors that policy makers must contemplate before settling on values for C_{FA}, $C_{M,}$ V_H, and V_{CR}. For one, false convictions take a financial toll: California and Illinois alone paid nearly half a billion dollars for false convictions overturned since 1989.[b] This estimate doesn't include the cost of incarceration, which is approximately $100 per day across the United States[c] and much higher in some locations.[d] Even more importantly, there are terrible social costs. A false conviction involves two tragic mistakes: the conviction of an innocent individual, and the continued freedom of the actual perpetrator. The Innocence Project reported that the actual perpetrators had been identified in 155 of the 358 DNA exoneration cases, and the actual perpetrators had been convicted of 150 additional violent crimes

[b] See https://www.innocenceproject.org/wrongful-convictions-cost-california-221-million; https://www.bettergov.org/news/wrongful-conviction-costs-keep-climbing.

[c] See https://www.federalregister.gov/documents/2018/04/30/2018-09062/annual-determination-of-average-cost-of-incarceration.

[d] See https://www.nytimes.com/2013/08/24/nyregion/citys-annual-cost-per-inmate-is-nearly-168000-study-says.html.

during their period of *wrongful liberty* (Baumgartner et al., 2018[e]). This number included Bobby Poole, accused of 20 additional crimes *after* he raped Jennifer Thompson and the second victim.[f]

Any policy decision regarding the determination of the optimal confidence level must take these financial and social costs into account. Only then can policy-makers determine the confidence level for the police to use that will distinguish identifications deemed to have probative value from those identifications that do not. Relying on identifications that meet or exceed the stipulated confidence level will result in a rate of false identifications that society judges acceptable. Some might argue that the occurrence of any false identifications is unacceptable, but the nature of the decision problem makes that goal unattainable. And keep in mind that not all false identifications result in false convictions. As Clark et al. (2015; see also Gould, Carrano, Leo, & Hail-Jares, 2013) point out, the impact of eyewitness evidence on the outcome of a trial depends on how identification evidence is utilized by a variety of legal actors, including the police, prosecuting and defense attorneys, judges (who rule on the admissibility of an identification), expert witnesses, and the members of a jury. But we can enhance the fairness of the criminal justice system by ensuring that these legal actors utilize accurate eyewitness evidence, fine-tuned by metacognitive proficiency, and filtered through existing and yet-to-be developed techniques that enable eyewitnesses to be reliable reporters of what they do and do not remember.

4. Conclusions

Memory is incomplete and prone to error (e.g., Schacter, 1999). When coupled with the frequent headlines proclaiming the exoneration of innocent people, it is not surprising that the general public and the criminal justice system deem memory evidence untrustworthy. However, this old science view of eyewitness memory must be revised, because memory reports carry metacognitive information that eyewitnesses can utilize to distinguish accurate from inaccurate identifications, and to differentiate among beneficial and misleading memory reports. The recent bridging of the historical divide between basic recognition memory research and applied research in eyewitness memory has provided a new narrative, shaped by

[e] See https://www.innocenceproject.org/dna-exonerations-in-the-united-states.

[f] See https://www.themarshallproject.org/2018/03/21/when-the-innocent-go-to-prison-how-many-guilty-go-free.

the adoption of three important traditions from basic memory research: signal-detection theory, the measurement of memory, and the contribution of metacognition.

These traditions have facilitated a number of important developments, several of which are featured in this chapter. Signal-detection theory demonstrates to policy makers how to use the weight of the two types of errors inherent to eyewitness identification decisions—the false identifications of innocent suspects and the missed identifications of guilty suspects—to determine the confidence level at which identifications are deemed accurate. The adoption of alternative means of measurement (ROC analysis) naturally disentangles memory accuracy from response bias, and reveals the strong relationship between confidence and accuracy (calibration). The utilization of ROC analysis has recast prior reform efforts, and the use of calibration analysis revealed the critically important role that response confidence can play, if the confidence judgment arises from the initial test of memory, and involves a confidence judgment taken at the time of that initial test. Consideration of other aspects of metacognition reveals how eyewitnesses can adjust what they report in a manner that maximizes the accuracy and informativeness of the reported information. If the potential and means for memory contamination are understood and controlled, the new science of eyewitness memory reveals that eyewitnesses are in control of their memories, rather than unwitting victims of a flawed memory system that they do not understand and cannot control.

The new science of eyewitness memory offers a very different perspective on Jennifer Thompson's identification of Ronald Cotton. In Jennifer's initial test (a photo lineup), she reported "I think this is the guy," only raising her confidence level as the detective pressed her for greater certainty. The district attorney requested a second test (a live lineup), with only Cotton being tested again. Again, she took several minutes before stating that Cotton "looks the most like him," with Jennifer indicating that she was certain only *after* the detective asked her "if she was certain." And the jury, instead of hearing Jennifer's initial report, heard her state in court that she was "absolutely sure" in her identification (from Garrett, 2011b).

According to the new science of eyewitness memory, and the research that we have presented in this chapter, how should an identification like Jennifer Thompson's unfold today? The police will test her memory only once, in a double-blind manner, accompanied by instructions indicating that the perpetrator may or may not be present in the lineup. Jennifer will make a confidence statement at the time of the initial identification, and that is the confidence statement on which the police, the attorneys, the judge, and the jury will rely. If Jennifer's level of confidence fails to reach the predetermined level, it will be treated as a non-identification. The old science view of eyewitness memory reinforces exactly the wrong message regarding Jennifer Thompson's incorrect identification. Rather than concluding that eyewitnesses can never be trusted, it was by way of her slow, unconfident identifications of Cotton, that Jennifer signaled that there was a good chance that she was making an error. It is our contention that false convictions like the one that ensnared Ronald Cotton can be decreased through the proper handling of eyewitness evidence, and a more comprehensive understanding of eyewitness memory.

References

Amendola, K. L., & Wixted, J. T. (2015). Comparing the diagnostic accuracy of suspect identifications made by actual eyewitnesses from simultaneous and sequential lineups in a randomized field trial. *Journal of Experimental Criminology*, *11*, 263–284.

Andersen, S. M., Carlson, C. A., Carlson, M., & Gronlund, S. D. (2014). Individual differences predict eyewitness identification performance. *Personality and Individual Differences*, *60*, 36–40.

Anderson, R. C., & Pichert, J. W. (1978). Recall of previously unrecalled information following a shift in perspective. *Journal of Verbal Learning and Verbal Behavior*, *17*, 1–12.

Arnold, G. F. (1906). *Psychology applied to legal evidence and other constructions of law*. Calcutta: Thacker, Spink & Co.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, *74*, 81–98.

Barnes, A. E., Nelson, T. O., Dunlosky, J., Mazzoni, G., & Narens, L. (1999). An integrative system of metamemory components involved in retrieval. In *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 287–313). Cambridge, MA: MIT Press.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. London: Cambridge University Press.

Baumgartner, F. R., Grigg, A., Ramírez, R., & Lucy, J. S. (2018). The mayhem of wrongful liberty: Documenting the crimes of true perpetrators in cases of wrongful incarceration. *Albany Law Review*, *81*(4).

Benjamin, A. S. (2007). Memory is more than just remembering: Strategic control of encoding, accessing memory, and making decisions. In A. S. Benjamin & B. H. Ross (Eds.), *The psychology of learning and motivation: Skill and strategy in memory use: Vol. 48.* (pp. 175–223). London: Academic Press.

Benjamin, A. S., Diaz, M. L., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review*, *116*(1), 84–115.

Benjamin, A. S., Tullis, J. G., & Lee, J. H. (2013). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1601–1608.

Bennett, S., Benjamin, A. S., Mistry, P. K., & Steyvers, M. (2018). Making a wiser crowd: Benefits of individual metacognitive control on crowd performance. *Computational Brain & Behavior*, *1*, 90–99.

Blackstone, W. (1765). *Commentaries on the laws of England: Vol. II,* Book IV. England, Oxford: Clarendon Press.

Bornstein, B. H., & Penrod, S. D. (2008). Hugo who? G. F. Arnold's alternative early approach to psychology and law. *Applied Cognitive Psychology*, *22*, 759–768.

Brewer, N., Weber, N., Wootton, D., & Lindsay, D. S. (2012). Identifying the bad guy in a lineup using confidence judgments under deadline pressure. *Psychological Science*, *23*, 1208–1214.

Brown-Schmidt, S., & Benjamin, A. S. (2018). How we remember conversations: Implications in legal settings. *Policy Insights From the Behavioral and Brain Sciences*, *5*, 187–194.

Carlson, C. A., & Carlson, M. A. (2014). An evaluation of perpetrator distinctiveness, weapon presence, and lineup presentation using ROC analysis. *Journal of Applied Research in Memory and Cognition*, *3*, 45–53.

Chan, J. C., & Lapaglia, J. A. (2011). The dark side of testing memory: Repeated retrieval can enhance eyewitness suggestibility. *Journal of Experimental Psychology: Applied*, *17*, 418–432.

Charman, S. D., Wells, G. L., & Joy, S. W. (2011). The dud effect: Adding highly dissimilar fillers increases confidence in lineup identifications. *Law and Human Behavior*, *25*, 479–500.

Clark, S. E. (2012). Costs and benefits of eyewitness identification reform: Psychological science and public policy. *Perspectives on Psychological Science*, 7, 238–259.

Clark, S. E., Benjamin, A. S., Wixted, J. T., Mickes, L., & Gronlund, S. D. (2015). Eyewitness identification and the accuracy of the criminal justice system. *Policy Insights From the Behavioral and Brain Sciences*, *2*, 175–186.

Clark, S. E., Brower, G. L., Rosenthal, R., Hicks, J. M., & Moreland, M. B. (2013). Lineup administrator influences on eyewitness identification and eyewitness confidence. *Journal of Applied Research in Memory and Cognition*, *2*, 158–165.

Clark, S. E., Moreland, M. B., & Gronlund, S. D. (2014). Evolution of the empirical and theoretical foundations of eyewitness identification reform. *Psychonomic Bulletin & Review*, *21*, 251–267.

Colloff, M., Wade, K., & Strange, D. (2016). Unfair lineups don't just make witnesses more willing to choose the suspect, they also make them more likely to confuse innocent and guilty suspects. *Psychological Science*, *27*, 1227–1239.

Crovitz, H. F., & Schiffman, H. (1974). Frequency of episodic memories as a function of their age. *Bulletin of the Psychonomic Society*, *4*, 517–518.

Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior*, *4*, 243–260.

Deffenbacher, K. A. (2008). Estimating the impact of estimator variables on eyewitness identification: A fruitful marriage of practical problem solving and psychological theorizing. *Applied Cognitive Psychology*, *22*, 815–826.

Dobolyi, D. G., & Dodson, C. S. (2013). Eyewitness confidence in simultaneous and sequential lineups: A criterion shift account for sequential mistaken identification overconfidence. *Journal of Experimental Psychology: Applied*, *19*, 345–357.

Dunning, D., & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identifications via inquiries about decision processes. *Journal of Personality and Social Psychology*, *67*, 818–835.

Egan, J. P. (1958). *Recognition memory and the operating characteristic. (Tech Note AFCRC-TN-58-51).* Bloomington, IN: Indiana University, Hearing and Communication Laboratory.

Egan, J. P. (1975). *Signal detection theory and ROC analysis.* New York: Academic Press.

Evans, J. R., & Fisher, R. P. (2011). Eyewitness memory: Balancing the accuracy, precision and quantity of information through metacognitive monitoring and control. *Applied Cognitive Psychology*, *25*, 501–508.

Fechner, G. T. (1966). In E. G. Boring & D. H. Howes (Eds.), *Elements of psychophysics (Vol. I)*. New York: Holt, Rinehart, & Winston (H. E. Adler, Trans.). (Original work published in 1860).

Fiechter, J. L., Benjamin, A. S., & Unsworth, N. (2016). The metacognitive foundations of effective remembering. In J. Dunlosky & S. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 307–324). New York, NY: Oxford University Press.

Fisher, R. P., Geiselman, R. E., & Amador, M. (1989). Field test of the cognitive interview: Enhancing the recollection of actual victims and interviewees of crime. *Journal of Applied Psychology*, *74*, 722–727.

Garrett, B. (2011a). *Convicting the innocent: Where criminal prosecutions go wrong*. Cambridge, MA: Harvard University Press.

Garrett, B. (2011b). *Getting in wrong: Convicting the innocent*. Retrieved June 21, 2018. http://www.slate.com/articles/news_and_politics/jurisprudence/features/2011/getting_it_wrong_convicting_the_innocent/how_eyewitnesses_can_send_innocents_to_jail.html.

Garry, M., Manning, C. G., Loftus, E. F., & Sherman, S. J. (1996). Imagination inflation: Imagining a childhood event inflates confidence that it occurred. *Psychonomic Bulletin & Review*, *3*, 208–214.

Garry, M., & Polaschek, D. L. L. (2000). Imagination and memory. *Current Directions in Psychological Science*, *9*, 6–10.

Geiselman, R. E., Fisher, R. P., Firstenberg, I., Hutton, L. A., Sullivan, S. J., Avetissian, I. V., et al. (1984). Enhancement of eyewitness memory: An empirical evaluation of the cognitive interview. *Journal of Police Science and Administration*, *12*, 74–80.

Glanzer, M., Adams, J., Iverson, G., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, *100*, 546–567.

Glucksberg, S., & McCloskey, M. (1981). Decisions about ignorance: Knowing that you don't know. *Journal of Experimental Psychology: Human Learning and Memory*, 7, 311–325.

Goldsmith, M., Koriat, A., & Weinberg-Eliezer, A. (2002). Strategic regulation of grain size memory reporting. *Journal of Experimental Psychology: General*, *131*, 73.

Gould, J. B., Carrano, J., Leo, R. A., & Hail-Jares, K. (2013). Predicting erroneous convictions. *Iowa Law Review*, *99*, 471–522.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Oxford, England: Wiley.

Gregory, W. L., Mowen, J. C., & Linder, D. E. (1978). Social psychology and plea bargaining: Applications, methodology, and theory. *Journal of Personality and Social Psychology*, *36*, 1521–1530.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). New York: Academic Press.

Gronlund, S. D., Carlson, C. A., Neuschatz, J. S., Goodsell, C. A., Wetmore, S., Wooten, A., et al. (2012). Showups versus lineups: An evaluation using ROC analysis. *Journal of Applied Research in Memory and Cognition*, *1*, 221–228.

Gronlund, S. D., Mickes, L., Wixted, J. T., & Clark, S. E. (2015). Conducting an eyewitness lineup: How the research got it wrong. In B. H. Ross (Ed.), *The psychology of learning and motivation: Vol. 63* (pp. 1–43). Waltham, MA: Academic Press.

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using ROC analysis. *Current Directions in Psychological Science*, *23*, 3–10.

Hasel, L. E., & Kassin, S. M. (2009). On the presumption of evidentiary independence: Can confessions corrupt eyewitness identifications? *Psychological Science*, *20*, 122–126.

Hellmann, D. F., & Memon, A. (2016). Attribution of crime motives biases eyewitnesses' memory and sentencing decisions. *Psychology, Crime & Law*, *22*, 957–976.

Horry, R., Brewer, N., & Weber, N. (2016). The grain-size lineup: A test of a novel eyewitness identification procedure. *Law and Human Behavior*, *40*, 147–158.

Hupbach, A., Gomez, R., & Nadel, L. (2009). Episodic memory reconsolidation: Updating or source confusion? *Memory*, *17*, 502–510.

Indow, T., & Togano, K. (1970). On retrieving sequence from long-term memory. *Psychological Review*, *77*, 317–331.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*, 3–28.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1304–1316.

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition*, *2*, 42–52.

Kassin, S. M., Ellsworth, P. C., & Smith, V. L. (1989). The "general acceptance" of psychological research on eyewitness testimony: A survey of the experts. *American Psychologist*, *44*, 1089–1098.

Kassin, S. M., Tubb, V. A., Hosch, H. M., & Memon, A. (2001). On the "general acceptance" of eyewitness testimony research: A new survey of the experts. *American Psychologist*, *56*, 405–416.

Key, K. N., Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D., Cash, D. K., & Lane, S. (2017). Lineup fairness affects postdictor validity and don't know responses. *Applied Cognitive Psychology*, *31*, 59–68.

Klobuchar, A., Steblay, N. K. M., & Caliguiri, H. L. (2006). Improving eyewitness identifications: Hennepin County's blind sequential lineup pilot project. *Cardozo Public Law, Policy, and Ethics Journal*, *2*, 381–414.

Köhnken, G., Milne, R., Memon, A., & Bull, R. (1999). A meta-analysis on the effects of the cognitive interview. *Psychology, Crime, & Law*, *5*, 3–27.

Koriat, A. (2018). When reality is out of focus: Can people tell whether their beliefs and judgments are correct or wrong? *Journal of Experimental Psychology: General*, *147*, 613–631.

Koriat, A., & Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: Distinguishing the accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology: General*, *123*, 297–315.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*, 490–517.

Krix, A. C., Sauerland, M., Lorei, C., & Rispens, I. (2015). Consistency across repeated eyewitness interviews: Contrasting police detectives' beliefs with actual eyewitness performance. *PLoS One*, *10*, e0118641.

Lampinen, J. M. (2016). ROC analyses in eyewitness identification research. *Journal of Applied Research in Memory and Cognition*, *5*, 21–33.

Lane, S. M., & Meissner, C. A. (2008). A "middle road" approach to bridging the basic-applied divide in eyewitness identification research. *Applied Cognitive Psychology*, *22*(6), 779–787.

Lehrer, J. (2010). *The truth wears off: Is there something wrong with the scientific method?* The New Yorker.

Lindsay, D. S. (1994). Memory source monitoring and eyewitness testimony. In D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Adult eyewitness testimony: Current trends and developments* (pp. 27–55). New York: Cambridge University Press.

Lindsay, R. C. L., Semmler, C., Weber, N., Brewer, N., & Lindsay, M. (2008). How variations in distance affect eyewitness reports and identification accuracy. *Law and Human Behavior*, *32*, 526–535.

Lindsay, R. C. L., & Wells, G. L. (1980). What price justice? Exploring the relationship of lineup fairness to identification accuracy. *Law and Human Behavior*, *4*, 303–313.

Lindsay, R. C. L., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology*, *70*, 556–564.

Link, S. W. (1994). Rediscovering the past: Gustav Fechner and signal detection theory. *Psychological Science*, *5*, 335–340.

Lockhart, R. S., & Murdock, B. B., Jr. (1970). Memory and the theory of signal detection. *Psychological Bulletin*, *74*, 100–109.

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology*, *7*, 560–572.

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, *12*, 361–366.

Loftus, E. F., & Kaufman, L. (1992). Why do traumatic experiences sometimes produce good memory (flashbulbs) and sometimes no memory (repression)? In E. Winograd & U. Neisser (Eds.), *Affect and accuracy in recall: Studies of "flashbulb" memories* (pp. 212–223). New York, NY: Cambridge University Press.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, *13*, 585–589.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). . Mahwah, NJ: Erlbaum.

Malpass, R. S., & Devine, P. G. (1981). Eyewitness identification: Lineup instructions and the absence of the offender. *Journal of Applied Psychology*, *66*, 482–489.

Manson v. Braithwaite. (1977). 432 U.S. 98.

McNicol, D. (1972). *A primer of signal detection theory*. London: Allen and Unwin.

Meisters, J., Diedenhofen, B., & Musch, J. (2018). Eyewitness identification in simultaneous and sequential lineups: An investigation of position effects using receiver operating characteristics. *Memory*, *26*, 1297–1309.

Memon, A., Meissner, C. A., & Fraser, J. (2010). The cognitive interview: A meta-analytic review and study space analysis of the past 25 years. *Psychology, Public Policy, and Law*, *16*, 340–372.

Mickes, L., Clark, S. E., & Gronlund, S. D. (2017). Distilling the confidence-accuracy message. *Psychological Science in the Public Interest*, *18*, 6–9.

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous and sequential lineups. *Journal of Experimental Psychology: Applied*, *18*, 361–376.

Mickes, L., & Gronlund, S. D. (2017). Eyewitness identification. In J. H. Byrne (Ed.), *Learning and memory: A comprehensive reference 2E, volume cognitive psychology of memory*. Elsevier.

Mickes, L., Seale-Carlisle, T. M., Wetmore, S. A., Gronlund, S. D., Clark, S. E., Carlson, C. A., et al. (2017). ROCs in eyewitness identification: Instructions vs. confidence ratings. *Applied Cognitive Psychology*, *31*, 467–477.

Münsterberg, H. (1908). *On the witness stand*. New York: Doubleday.

National Research Council. (2014). *Identifying the culprit: Assessing eyewitness identification*. Washington, DC: National Academies Press.

Neil v. Biggers. (1972). 409 U.S. 188.

Neisser, U. (1981). John Dean's memory: A case study. *Cognition*, *9*, 1–22.

Neuschatz, J. S., Wetmore, S. A., Key, K., Cash, D., Gronlund, S. D., & Goodsell, C. A. (2016). Comprehensive evaluation of showups. In M. Miller & B. Bornstein (Eds.), *Advances in psychology and law*. New York: Springer.

Odinot, G., Wolters, G., & Lavender, T. (2009). Repeated partial eyewitness questioning causes confidence inflation but not retrieval-induced forgetting. *Applied Cognitive Psychology*, *23*, 90–97.

Oeberst, A. (2015). How good are future lawyers in judging the accuracy of reminiscent details? The estimation-observation gap in real eyewitness accounts. *The European Journal of Psychology Applied to Legal Context*, *7*, 73–79.

Osth, A. F., Dennis, S., & Heathcote, A. (2017). Likelihood ratio sequential sampling models of recognition memory. *Cognitive Psychology*, *92*, 101–126.

PBS Online. (2018). *Frontline: What Jennifer saw*. Retrieved June 21, 2018, from http://www.pbs.org/wgbh/pages/frontline/shows/dna/interviews/thompson.html.

Penrod, S., & Cutler, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, and Law*, *1*, 817–845.

Pezdek, K. (1977). Cross-modality semantic integration of sentence and picture memory. *Journal of Experimental Psychology: Human Learning and Memory*, *3*, 515–524.

Pezdek, K., Finger, K., & Hodge, D. (1997). Planting false childhood memories: The role of event plausibility. *Psychological Science*, *8*, 437–441.

Pozzulo, J. D., & Lindsay, R. C. L. (1999). Elimination lineups: An improved identification procedure for child eyewitnesses. *Journal of Applied Psychology*, *84*, 167–176.

Rakoff, J. S. (2014). *Why innocent people plead guilty*. The New York Review of Books.

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*, 518–535.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.

Roediger, H. L., & McDermott, K. B. (2000). Distortions of memory. In F. I. M. Craik & E. Tulving (Eds.), *The Oxford handbook of memory* (pp. 149–164). Oxford, England: Oxford University Press.

Rotello, C. M. (2017). Signal detection theories of recognition memory. In J. T. Wixted (Ed.), *Cognitive psychology of memory: Vol. 4. Learning and memory: A comprehensive reference (2nd ed.). Elsevier.*

Rotello, C. M., & Chen, T. (2016). ROC curve analyses of eyewitness identification decisions: An analysis of the recent debate. *Cognitive Research: Principles and Implications*, *1*, 1–12.

Rubin, D. C. (1982). On the retention function of autobiographical memory. *Journal of Verbal Learning and Verbal Behavior*, *21*, 21–38.

Sauer, J. D., Brewer, N., & Weber, N. (2008). Multiple confidence estimates as indices of eyewitness memory. *Journal of Experimental Psychology: General*, *137*, 528–547.

Schacter, D. L. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. *American Psychologist*, *54*, 182–203.

Seale-Carlisle, T. M., & Mickes, L. (2016). US lineups outperform UK lineups. *Royal Society Open Science*, *3*, 1–9.

Semmler, C., Dunn, J., Mickes, L., & Wixted, J. T. (2018). The role of estimator variables in eyewitness identification. *Journal of Experimental Psychology. Applied*, *24*, 400–415.

Shaw, J. S., Bjork, R. A., & Handal, A. (1995). Retrieval-induced forgetting in an eyewitness-memory paradigm. *Psychonomic Bulletin & Review*, *2*, 249–253.

Shaw, J. S., III, & McClure, K. A. (1996). Repeated postevent questioning can lead to elevated levels of eyewitness confidence. *Law and Human Behavior*, *20*, 629–654.

Simons, D. J., & Chabris, C. F. (2011). What people believe about how memory works: A representative survey of the U.S. population. *PLoS One*, *6*(8), e22757.

Smith, A. M., Wells, G. L., Lindsay, R. C. L., & Penrod, S. D. (2017). Fair lineups are better than biased lineups and showups, but not because they increase underlying discriminability. *Law and Human Behavior*, *41*, 127–145.

Sporer, S. L. (1992). Post-dicting eyewitness accuracy: Confidence, decision-times and person descriptions of choosers and non-choosers. *European Journal of Social Psychology*, *22*, 157–180.

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence–accuracy relation in eyewitness identification studies. *Psychological Bulletin*, *118*, 315–327.

Stanley, S. E., & Benjamin, A. S. (2016). That's not what you said the first time: A theoretical account of the relationship between consistency and accuracy of recall. *Cognitive Research: Principles and Implications*, *1*, 14.

Steblay, N. K., & Dysart, J. E. (2016). Repeated eyewitness identification procedures with the same suspect. *Journal of Applied Research in Memory and Cognition*, *5*, 284–289.

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *24*, 1397–1410.

Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin*, *99*, 181–188.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–26.

Thompson-Cannino, J., Cotton, R., & Torneo, E. (2010). *Picking Cotton: Our memoir of injustice and redemption*. New York, NY: St. Martin's Press.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*, 359–380.

Weber, N., & Brewer, N. (2008). Eyewitness recall: Regulation of grain size and the role of confidence. *Journal of Experimental Psychology: Applied*, *14*, 50–60.

Weber, N., & Perfect, T. J. (2012). Improving eyewitness identification accuracy by screening out those who say they don't know. *Law and Human Behavior*, *36*, 28–36.

Wells, G. L. (1978). Applied eyewitness-testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology*, (12), 1546–1557.

Wells, G. L. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology*, *14*, 89–103.

Wells, G. L. (1993). What do we know about eyewitness identification? *American Psychologist*, *48*, 553–571.

Wells, G. L., & Bradfield, A. L. (1998). "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology*, *83*, 360–376.

Wells, G. L., & Bradfield, A. L. (1999). Eyewitnesses' recollections of their certainty, witnessing conditions, and identification decisions: The distorting effects of feedback. *Psychological Science*, *10*, 138–144.

Wells, G. L., & Lindsay, R. C. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin*, *88*(3), 776–784.

Wells, G. L., Memon, A., & Penrod, S. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest*, *7*, 45–75.

Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology*, *54*, 277–295.

Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology*, *78*, 835–844.

Wells, G. L., Smalarz, L., & Smith, A. M. (2015). ROC analysis of line-ups does not measure underlying discriminability and has limited value. *Journal of Applied Research in Memory and Cognition*, *4*, 313–317.

Wells, G. L., Steblay, N. K., & Dysart, J. E. (2012). Eyewitness identification reforms: Are suggestiveness-induced hits and guesses true hits? *Perspectives on Psychological Science*, *7*, 264–271.

Wetmore, S. A., McAdoo, R. M., Gronlund, S. D., & Neuschatz, J. S. (2017). Impact of fillers on lineup performance. *Cognitive Research: Principles and Implications*, *2*, 48. 1–13.

Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon probative value and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science*, *7*, 275–278.

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review*, *121*, 262–276.

Wixted, J. T., & Mickes, L. (2018). Recognition memory in the laboratory and in the real world. In *Keynote address presented at the International Meeting of the Psychonomic Society, Amsterdam, The Netherlands*.

Wixted, J. T., Mickes, L., Clark, S., Gronlund, S., & Roediger, H. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist*, *70*, 515–526.

Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E., & Wells, W. (2016). The reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 304–309.

Wixted, J. T., Mickes, L., Wetmore, S. A., Gronlund, S. D., & Neuschatz, J. S. (2017). ROC analysis in theory and practice. *Journal of Applied Research in Memory and Cognition*, *6*, 343–351.

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, *18*, 10–65.

Yates, S. Q. (2017). *Memorandum for heads of department law enforcement components*. https://www.justice.gov/file/923201/download. Retrieved August, 30, 2018.

Zaragoza, M. S., & Lane, S. M. (1994). Sources of misattribution and suggestibility of eye witness testimony. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 934–945.